Comparing Human and Machine Generated Text for Sentiment

WingYin Ha and Diarmuid P. O'Donoghue[a]
Department of Computer Science, Maynooth University, Co. Kildare, Ireland
wingyin.ha.2019@mumail.ie, diarmuid.odonoghue@mu.ie

Keywords: Large Language Model (LLM), Parallel Corpus, Sentiment, Human Machine Comparison, Evaluation.

Abstract: This paper compares human and machine generated texts, focusing on a comparison of their sentiment. We use two corpora; the first being the HC3 question and answer texts. We present a second corpus focused on human written text-materials sourced from psychology experiments and we used a language model to generate stories analogous to the presented information. Two sentiment analysis tools generated sentiment results, showing that there was a frequent occurrence of statistically significant differences between the sentiment scores on the individual sub-collections within these corpora. Generally speaking, machine generated text tended to have a slightly more positive sentiment than the human authored equivalent. However, we also found low levels of agreement between the Vader and TextBlob sentiment-analysis systems used. Any proposed use of LLM generated content in the place of retrieved information needs to carefully consider subtle differences between the two – and the implications these differences may have on down-stream tasks.

1 INTRODUCTION

The abilities of Large Language Models (LLM) like ChatGPT are still poorly understood and greater understanding is essential in the face of widespread adoption, to ensure safe and reliable utilization. This paper uses two parallel corpora of human and machine originated text to find any similarities and notable differences between them. This paper focuses on the sentiment of these parallel texts, using five distinct parallel collections.

(Yiu *et al*, 2023) argue that LLM are cultural technologies that enhance cultural transmission. (Connell and Lynott, 2024) discussed the strengths and weaknesses of large language models to foster better understanding of human cognition. (Gibney, 2024) note that statements written in the African American English (AAE) dialect (widely spoken in the United States) have revealed strong racial biases in ChatGPT, making it more likely to associate fictionalised speakers with less-prestigious jobs and even more like to recommend the death penalty for a fictional defendant. (Mitchell, 2021) critiqued the ability of LLM to form concepts, abstractions and even make analogies.

Some application may consider machine generated texts as an alternative to retrieving text from a corpus. But this approach assumes equivalence between generated and human when text. This putative equivalence is put to the test in this paper by analysing two corpora of aligned human and machine generated texts. This work also contributes to ongoing work on model collapse (Feng *et al*, 2024) and the impact of machine generated data in training LLM.

This paper evaluates one pre-existing corpus and presents a novel corpus composed of analogous story pairs. We shall argue that these generated analogous stories offer a better mechanism to explore the innate bias contained within LLM.

We use existing technologies to investigate the output of LLM for any sentiment bias and differences between human and LLM originated text. For a comparison of Vader and TextBlob for sentiment analysis see (Bonta *et al*, 2019).

There has been a sigificant amount of recent work on comparing human and LLM generated text. (Katib *et al*, 2023; Liao et al, 2023), with some of this work focusing on detecting machine generated text in the context of plagiarism (Khalil and Erkan, 2023; Cotton *et al*, 2024). This paper differs from previous work in several regards. Firstly, this paper focuses on comparing the sentiment of texts. Secondly, we are not aware of any revious work on using the analogy

[a] https://orcid.org/0000-0002-3680-4217

approach to generate text, leading to the Analogy Materials Corpus (AMC) used in this paper. Finally, we compare the human and machine AMC texts using sentiment.

This paper is structured as follows. Firstly we discuss the background for comparing human and LLM generated text. We describe HC3 corpus (and each of its constituent sub-collections), before describing how the human portion of AMC coprus was compiled. We then detail how an LLM was used to generate analogous texts before presenting an analysis of the AMC texts.

We briefly describe our system before presenting and analysing our resutls on the HC3 corpus before analysing the AMC results. Finaly some conclusions and future work are discussed.

2 BACKGROUND

Widespread adoption of LLM since ChatGPT has raised concerns about its output and the presence of any hidden biases therein. Studies of LLM have shown the larger and more powerful models possess some surprising abilities, such as the ability to interpret analogical comparisons (Webb *et al*, 2023), they have shown an ability in terms of Theory of Mind (Strachan *et al*, 2024). (Ichen & Holyoak, 2024) evaluated text that they were confident was not included in any LLM training data to evaluate GPT-4's ability to detect & explain any contained metaphors. We did not follow this effort to ensure the novelty of the query text as we wish to better reflect typical usage of these LLM, which includes a combination of novel and familiar text in each query.

We argue that analogies offer a better mechanism to explore the sentiment of machine generated text. While the question-and-answer scenario restricts the range of possible responses to a prompt, generating novel analogies in contrast opens a much wider range of response types and topics. The semantic restriction that questions imposed on the range of possible answers is in effect removed by requesting the LLM to generate a comparable story both one that requires, or is even founded upon, a reasonable semantic distanced between the original and generated stories.

Thus, we argue, that generating analogous stories to a presented text imposes fewer constraints on the responses and thereby uncovers a more faithful reflection of the contents and biases contained within the LLM machine itself. Later in this paper we shall detail the Analogy Materials Corpus (AMC) that contains parallel human and machine generated text, containing analogous pairs of English texts.

3 PARALLEL CORPORA OF HUMAN AND MACHINE TEXT

This section describes two corpora of parallel human and machine generated text. Firstly, the pre-existing the Human ChatGPT Comparison Corpus (HC3) (Guo *et al*, 2023), which was produced under a question-answer scenario, by recording comparable answers to a given list of questions.

3.1 HC3 Corpus

The Human ChatGPT Comparison Corpus (HC3) is a collection of we collected 24,322 questions, 58,546 human answers and 26,903 ChatGPT answers (Guo *et al*, 2023). The corpus contains paired responses from both human experts and ChatGPT, allowing comparison of broad trends in the ability to each to generate text. Questions were grouped according to theme, including; open-domain, financial, medical, legal, and psychological areas. Their lexical analysis showed that ChatGPT uses more NOUN, VERB, DET, ADJ, AUX, CCONJ and PART words, while using less ADV and PUNCT words.

Sentiment analysis of text in (Guo *et al*, 2023) used a version of Roberta that was fine-tuned on a Twitter corpus. Additionally, their sentiment analysis focused on the collection as a whole and didn't examine the individual sub-collections. Limitations of the previous work include difficulty in reproducing the results (because of fine-tuning) and difficulty in benchmarking results against more established sentiment analysis models. Our sentiment analysis uses Vader (Hutto and Gilbert, 2014) one of the most widely used sentiment analysis models. This is compared with the newer TextBlob (Loria, 2018) model.

Table 1: Word count on the HC3 texts.

	Medicine		Finance		Open_qa		Wiki_csai	
	H	G	H	G	H	G	H	G
M	82	196	176	233	31	356	193	183
SD	46	76	160	100	19	161	124	49

The HC3_medical group contains text with strongly positive and strongly negative sentiments while the finance collection is dominated by a neutral sentiment.

The human texts contained an average of 120.5 words while the GPT texts were approximately twice that length at 241.3 words.

4 ANALOGY MATERIALS CORPUS (AMC)

Abgaz *et al* (2017) examined the characteristics of analogies between the text of publications in computer graphics. (O'Donoghue *et al*, 2015) showed how analogies can help stimulate creative thinking. Mitchell (2023) argues that large language models do not properly match human ability to form abstractions and use analogies. But recent studies (Webb *et al*, 2023) have shown that the bigger LLM models like ChatGPT possess the ability to correctly interpret analogical comparisons, including those between text stories.

In this paper we used an LLM to generate stories that are intended to be analogous to presented (human authored) stories.

As stated earlier, we see the generation of analogies as a powerful mechanism for evaluating the preferences and biases in LLM. Unlike the Question answer scenario that constrains the topic and arguably biases the expression of an answer, the hallmark of analogy is the presence a noticeable semantic difference between the presented information and its newly created analogous version.

The Structure Mapping Theory (Gentner, 1983) of analogy identifies the hallmarks of analogy as a semantic difference coupled with identifiable parallel systems of information between the two analogous scenarios.

We created the Analogy Materials Corpus [2] (AMC), composed of 169 short text stories selected from almost 40 distinct publications reporting empirical cognitive studies, including those reported by (Webb *et al*, 2023). These were first written by analogy researchers who were exploring the factors influence the human ability to interpret analogical comparisons and these materials (in the form of pairs of texts) were subsequently used on human experimental participants. Some of these experiments presented a target problem with alternate sources, to ascertain conditions that induce the expected solution in subjects. Other experiments couple a source solution with alternate target problems to see which are solved. Different participant groups are given different materials with solutions rates being studied, to ascertain different factors impacting on the analogy process. These working memory factors and the order of presentation of information (Keane, 1997) to the role of related sources on inducing general rules and their impact on subsequent reasoning (Gick and Holyoak, 1983).

[2] https://www.kaggle.com/diarmuidodonoghue/datasets

Stories were selected from within these materials and Llama2 was used to generate novel source stories that were analogous to each presented text. The 7bn parameter version was used with the *temperature* set to Zero (for reproducible results) but other LLM parameters were generally left with default values. This produced parallel corpus of human and machine authored texts and this paper treats the pre-existing human written sources and the machine generated stories as a kind of parallel corpus.

For reproducibility, the *temperature* parameter was set to 0. Initial testing indicated that best results were produced by setting: *role* to *user* and *model* was set to *instruct*. Responses from Llama2 were frequently accompanied by standard pre-pended text such as *"Sure. Here is a story that is analogous to the given story:".* Because these statements were not related specifically to the query and because they appeared in many answers, they were removed from each machine generated output.

4.1 Word Count and Vocabulary Size

The corpus contains 338 distinct text stories, in two paired collections of 169 texts each. The human texts had an average of 254.2 (SD=96.9) words, ranging in size from 67 to 882 words. The machine texts had an average size of 160.0 (SD=100.1), ranging in size from 17 to 523 words.

The average number of unique words in the human texts was 89.1 (SD=44.9) ranging from 16 to 233 unique words. The machine generated texts averaged 126.7 (SD=30.1) words, ranging in size from 49 to 228 distinct words.

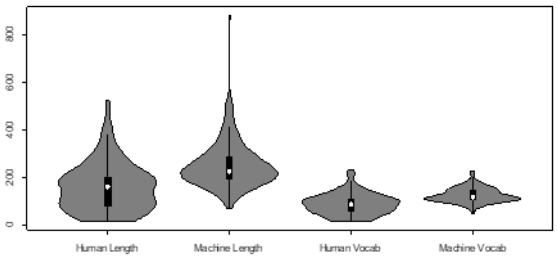

Figure 1: Machine generated texts were longer than the human texts (left) and used a larger vocabulary than the human texts.

We conclude that the LLM generated stories are highly comparable in size to the presented text. Furthermore, an *ad hoc* analysis of the generated stories suggested the presence of a semantic

difference between the original and machine generated texts.

4.2 Part of Speech Analysis

We also performed a lexical analysis on the human and machine produced texts, as was performed in (Guo *et al*, 2003). We focus our analysis on the main lexical categories of; noun, verb, pronoun, adpositions as these are some of the lexical categories of greatest relevance to interpreting analogical comparisons. NLTK was used to perform the lexical analysis.

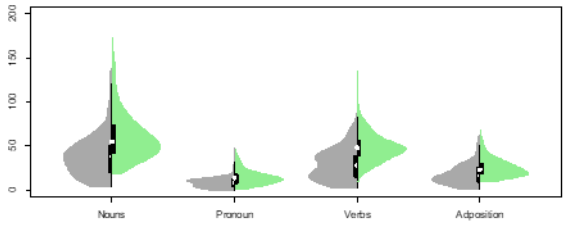

Figure 2: Lexical comparison of the human (grey) and machine (green) generated texts from the Analogy Materials Corpus.

Figure 2 depicts the number of words in each of the four analyzed lexical categories. The first split violin plot quantifies the number of nouns contained in each text. The left side of each violin (grey) quantifies the number of words in each text written by a human, while the right side of each violin (green) depicts the number of nouns in each of the machine generated texts.

These results show a high degree of similarity between the number of words in each of these categories. In this paper it was not necessary to validate whether the original and machine generated texts were in fact truly analogous to one another – or if they were merely similar to one another in some unspecified but abstract way.

4.3 Pairwise Differences

We performed a more detailed pairwise comparison of the difference in size between the human and machine originated text, with the results summarised in the table below. These again reflect the fact that machine generated texts are slightly larger than the corresponding human texts.

Table 2 details the size differences between the paired texts for each of the examined lexical categories, as well as for total number of words.

Overall we find that the machine generated texts are longer than the human texts.

Table 2: Differences between the human and ACM texts.

	Words	**Adp**	**Noun**	**Pron**	**Verb**
Mean	-94.1	-7.9	-23.9	-6.1	-19.8
SD	119.8	14.5	34.7	10.0	23.0

5 SYSTEM DESCRIPTION

A system was written in Python version 3.9.13 to determine the sentiment scores for each individual text in the HC3 corpus and in the AMC corpus. This system used the libraries; NLTK (Natural Language Tool Kit) 3.8.1 and its *sentiment.vader* library and TextBlob (*PatternAnalyser*) version 0.18. All experiments were performed on a standard laptop computer and all execution times were in the order of seconds and are not reported further.

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool. TextBlob is a general-purpose text progressing system that includes a sentiment analysis system.

Vader returns scores ranging of -1 for the most negative sentiment and +1 for the most positive. Similariy, TextBlob returns polarity scores also in the range from -1 to 1. The following results were produced by these two systems. The two systems occasionally reveal the same insight into human and machine generated text, put frequently the two systems give somewhat different insights.

6 RESULTS AND ANALYSIS

We now present the results and analysis of the sentiment analysis of the two corpora. We begin with the HC3 results and each of its constituent collections, followed by the AMC analogy results.

6.1 HC3 - Overall

Vader (Hutto and Gilbert, 2014) and TextBlob (Loria, 2018) showed very different results on the HC3-medicine corpus. Vader identified stronger positivity on three machine generated texts, but humans showed more positivity on the other medicine collection.

Table 3 details sentiment scores for the 4 collections in HC3, with H indicating human text and G for machine generated text.

Table 3: Average Vader Sentiment on the HC3 texts.

	Medicine		Finance		Open_qa		Wiki_csai	
	H	G	H	G	H	G	H	G
Mean	.34	.08	.46	.74	.11	.38	.49	.67
SDev	.60	.79	.57	.45	.40	.55	.54	.47

Table 4 shows sentiment values generated by TextBlob for each of the HC3 collections. TextBlob scores were more neutral than the corresponding Vader scores.

Table 4: Average TextBlob Sentiment on the HC3 Corpus.

	Medicine		Finance		Open_qa		Wiki_csai	
	H	G	H	G	H	G	H	G
Mean	.14	.12	.10	.12	.05	.09	.05	.07
SDev	.46	.53	.44	.45	.23	.41	.40	.45

6.1.1 Levels of Agreement

We compared the sentiment of human and machine text for the HC3 corpus, dividing the resulting differences into three categories, as follows:

Strong Disagreement: difference > 0.5
Disagreement: <= 0.1 difference <= 0.5
Strong Agreement: difference <= 0.1

This categorisation divided the overall HC3 corpus into three approximately equally sized categories, accounting for between 31% and 35% of the overall corpus in each of the three categories.

Table 5: Sentiment comparison between human and machine text using Vader.

	Medicine	Finance	Open_qa	Wiki_csai
Strong Disagree	55.05	36.49	43.15	31.21
Disagree	29.89	32.87	43.89	35.50
Strong Agree	15.06	30.64	18.98	33.38

Table 5 shows the greatest degree of dissimilarity between human and machine text in the Medicine collection, while Finance and Wiki_csai showed the greater levels of agreement in sentiment.

6.2 HC3 - Medicine

Looking more closely at HC3-medicine, Vader identified more positivity in human texts (M=0.34, SD=0.6) and more neutral sentiment in the GPT text M=0.08, but also showed machine had greater variation (SD=0.79). In contrast, TextBlob showed almost the opposite trend, with neutral scores dominating and few positive and negative scores for both human and machine text.

Figure 3 below shows the distribution of sentiment scores for these texts. Figure 3 contains two graphs; the left bar-graph depicts the Vader results while the right shows the TextBlob results. Within each graph the human results are depicted in blue while results produced on machine generated text are shown in red.

The Vader results show that human text had a larger number of neutral scores while the machine generated text had more highly negative and far more highly positive scores. The TextBlob results show a different pattern, with most scores centered on neutral sentiments. TextBlob generally showed a greater Sentiment was evident in the human written text.

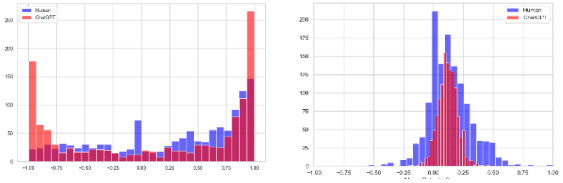

Figure 3: HC3_medicine scores of human (blue) and machine (red) generated text, using Vader (left) and TextBlob (right) polarity scores.

A Mann-Whitney analysis of the human and machine scores on HC3 - Medicine gave a two-tailed z-score of 1.07469. and the p-value is < 0.14. Thus, the difference in sentiment scores was not significant at $p < 0.1$.

6.3 HC3 – Finance

Figure 4 (and the subsequent diagrams in this section) also depict Vader results on the left and TextBlob on the right. Vader results show that ChatGPT text showed a far higher incidence of highly positive scores. The majority of both human and machine text showed that there were few texts with low levels of positive or with negative sentiment.

TextBlob scores indicate high levels of neutral or slightly positive sentiment on both human and machine text. However, the machine text seems to display very slightly more positive sentiment. Overall however, there appeared to be broad agreement under sentiment between human and machine generated text for this sub-collection.

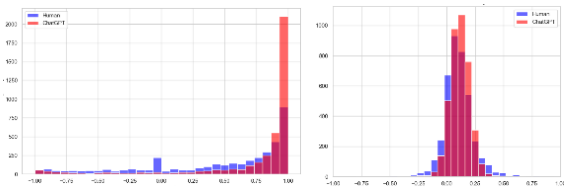

Figure 4: Vader and TextBlob found similar pattern of sentiment scores on HC3_Finance between the human and machine generated text.

A Mann-Whitney analysis of the human and machine scores on HC3_Finance gave a two-tailed z-score of -6.9308 and the p-value is < 0.0001. Thus, the difference was significant at $p < 0.1$. So, there is a statistically significant difference in the sentiment scores between these two collections.

6.4 HC3 – Open_qa

Vader results (Figure 5) show the human text was dominated by a neutral sentiment well the machine text was dominated by very positive sentiment. TextBlob analysis loosely echoed the dominance of neutral sentiment in the human text, while the sentiment of machine text was centered on a very slightly positive sentiment.

In this collection we see a moderate degree of agreement between Vader and TextBlob, both showing human texts to be predominantly neutral. However Vader detected a greater degree of positivity than Texblob.

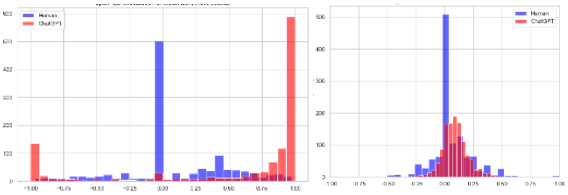

Figure 5: Comparing HC3_Open_qa scores on human and machine generated text, using Vader (left) and TextBlob (right).

A Mann-Whitney analysis of the human and machine scores on HC3_Wiki_open_qa gave a two-tailed z-score of -6.9308 and the p-value is < 0.0001. Thus, the difference was significant at $p < 0.1$. So, there is a statistically significant difference in the sentiment scores between these two collections.

6.5 HC3 – Wiki_csai

Figure 6 shows the sentiment analysis results on the wiki_csai collection from HC3. Vader scores the dominance of positive sentiment for both human and machine generated text. However the machine generated text exhibits a greater number of highly positive scores.

TextBlob analysis showed a similar trend between human and machine texts, centered on the dominance of very slightly positive scores. However human texts displayed a greater incidence of this sentiment then was found in the machine texts.

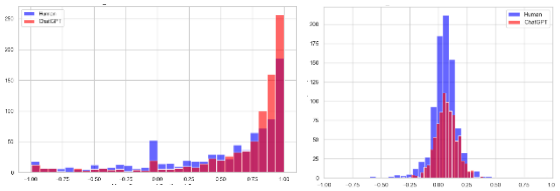

Figure 6: HC3_csai human and machine text scores, using Vader (left) and TextBlob (right).

A Mann-Whitney analysis of the human and machine scores on HC3_Wiki_csai gave a two-tailed z-score of -6.9308 and the p-value is < 0.0001. Thus, the difference was significant at $p < 0.1$. So, there is a statistically significant difference in the sentiment scores between these two collections.

6.6 Analogy Materials Corpus (AMC)

Finally, Vader analysis (Figure 7) of the AMC corpus revealed the machine generated text showed a large number of highly positive sentiments. While the human text showed a large positive sentiment, it also had a broader distribution of sentiment scores.

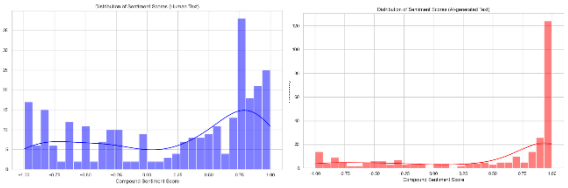

Figure 7: The machine generated AMC text showed a strong bias towards highly positive sentiment scores.

TextBlob analysis (Figure 8) indicated the human and Llama2 text both had a tendency towards a neutral sentiment. However, Llama2 text was slightly more positive than the original human texts.

While the sentiment in these results were far from identical it did show a surprising degree of agreement, given the very general nature of the task of generating analogous text. However this level of agreement should be seen in the light of the original texts being dominated by sentiment scores very close to 0 and

with these scores also displaying something akin to a normal distribution.

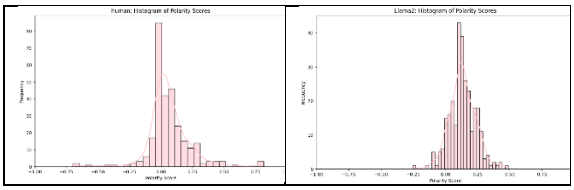

Figure 8: TextBlob polarity on human and machine AMC text.

A Mann-Whitney analysis of the human and machine scores on the AMC corpus gave a two-tailed z-score of -4.42117 and the p-value is < 0.0001. Thus, the difference was significant at $p < 0.1$. So, there is a statistically significant difference in the sentiment scores between these human and machine texts. This was an interesting result as generating source analogs was seen as giving a great deal of freedom to the LLM in terms of its chosen subject matter and the manner in which that was expressed.

7 CONCLUSIONS

We present a comparison between text written by humans with comparable machine generated text. The objective in this paper was to assess Large Language Machines, like ChatGPT and Llama2, for biases and significant differences that distinguish their output from human text. This paper focuses on sentiment of the text, using two established sentiment analysis systems, Vader and TextBlob, to perform the analysis.

Two corpora were used; firstly the existing Human ChatGPT Comparison Corpus (HC3) corpus, containing human and machine responses to questions. Secondly we present the Analogy Materials Corpus (AMC) containing human writes texts used in psychology experiments, with the Llama2 LLM being tasked with generating analogous texts to the presented stories.

Many instances of statistically significant differences between the sentiment of human and machine text were identified. In general, machine generated text seemed to exhibit a more positive sentiment than the comparable human text. These differences were often relatively small in magnitude, but the HC3_medicine collection showed the greatest difference in the pattern of sentiment scores.

Based on these findings we additionally conclude that any putative use of LLM generated content in the place of retrieved (human) information needs to carefully consider (the often subtle) differences between human and LLM generated content.

ACKNOWLEDGEMENTS

This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 21/FFP-P/10118. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

REFERENCES

Abgaz, Yalemisew; O'Donoghue, Diarmuid P.; Hurley, Donny; Chaudhry, Ehtzaz; Zhang, Jian Jun. Characteristics of Pro-c Analogies and Blends between Research Publications, *International Conference on Computational Creativity (ICCC)*, pp 1 – 8, Atlanta, GA, USA, June 2017.

Bonta, Venkateswarlu, Nandhini Kumaresh, and Naulegari Janardhan. "A comprehensive study on lexicon based approaches for sentiment analysis." *Asian Journal of Computer Science and Technology* 8, no. S2: 1-6. (2019).

Connell, Louise, and Dermot Lynott. "What Can Language Models Tell Us About Human Cognition?." *Current Directions in Psychological Science* 33, no. 3: 181-189. (2024).

Cotton, Debby RE, Peter A. Cotton, and J. Reuben Shipway. "Chatting and cheating: Ensuring academic integrity in the era of ChatGPT." *Innovations in Education and Teaching International* 61, no. 2 (2024): 228-239.

Feng, Yunzhen, Elvis Dohmatob, Pu Yang, Francois Charton, and Julia Kempe. "Beyond Model Collapse: Scaling Up with Synthesized Data Requires Reinforcement." *arXiv preprint arXiv:2406.07515* (2024).

Gentner, Dedre. "Structure-mapping: A theoretical framework for analogy." *Cognitive Science* 7, no. 2 (1983): 155-170.

Gibney, Elizabeth. "Chatbot AI makes racist judgements on the basis of dialect." *Nature* 627, no. 8004: 476-477. (2024).

Gick, Mary L., and Keith J. Holyoak. "Schema induction and analogical transfer.", *Cognitive Psychology*, 15, no. 1: 1-38 (1983).

Guo, Biyang, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. "How close is ChatGPT to human experts? comparison corpus, evaluation, and detection." *arXiv preprint arXiv:2301.07597* (2023).

Hutto, Clayton, and Eric Gilbert. “Vader: A parsimonious rule-based model for sentiment analysis of social media text.” In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216-225. (2014).

Ichien, Nicholas, Dušan Stamenković, and Keith Holyoak. “Interpretation of Novel Literary Metaphors by Humans and GPT-4.” In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 46. 2024.

Katib, Iyad, Fatmah Y. Assiri, Hesham A. Abdushkour, Diaa Hamed, and Mahmoud Ragab. “Differentiating chat generative pretrained transformer from humans: detecting ChatGPT-generated text and human text using machine learning.” *Mathematics* 11, no. 15 (2023): 3400.

Keane, Mark T. “What makes an analogy difficult? The effects of order and causal structure on analogical mapping.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23, no. 4 (1997): 946.

Khalil, Mohammad, and Erkan Er. “Will ChatGPT G et You Caught? Rethinking of Plagiarism Detection.” *In International Conference on Human-Computer Interaction*, pp. 475-487. Cham: Springer Nature Switzerland, 2023.

Liao, Wenxiong, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang et al. “Differentiating ChatGPT-generated and human-written medical texts: quantitative study.” *JMIR Medical Education* 9, no. 1 (2023): e48904.

Loria, Steven. “TextBlob Documentation.” *Release 0.15 2, no. 8*: 269. (2018)

Mitchell, Melanie. “Abstraction and analogy in AI.” *Annals of the New York Academy of Sciences* 1524, no. 1 (2023): 17-21. DOI: 10.1111/nyas.14995.

O'Donoghue, Diarmuid, Yalemisew Abgaz, Donny Hurley, and Francesco Ronzano. “Stimulating and simulating creativity with Dr Inventor.” *International Conference on Computational Creativity (ICCC)*, Park City, Utah, USA, pp220-227 (2015).

Strachan, James WA, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena *et al.* “Testing theory of mind in large language sorrymodels and humans.” *Nature Human Behaviour*: 1-11, (2024).

Webb, Taylor, Keith J. Holyoak, and Hongjing Lu. “Emergent analogical reasoning in large language models.” *Nature Human Behaviour* 7, no. 9: 1526-1541. (2023).

Yiu, Eunice, Eliza Kosoy, and Alison Gopnik. “Transmission versus truth, imitation versus innovation: What children can do that large language and language-and-vision models cannot (yet).” *Perspectives on Psychological Science* (2023): 17456916231201401.